Semiparametric Regression for the Social Sciences

Luke Keele
Ohio State University, U.S.A.
Semiparametric Regression for the Social Sciences
Semiparametric Regression for the Social Sciences

Luke Keele

Ohio State University, U.S.A.

John Wiley & Sons, Ltd
3.2.3  B-splines ................................... 58
3.2.4  Knot Placement and Numbers .................... 59
3.2.5  Comparing Spline Models ........................ 61
3.3  Splines and Overfitting............................... 62
3.3.1  Smoothing Splines ............................ 64
3.3.2  Splines as Mixed Models........................ 69
3.3.3  Final Notes on Smoothing Splines ................. 73
3.3.4  Thin Plate Splines. ............................. 74
3.3.5  Comparing Spline Models ....................... 61
3.4  Inference for Splines ................................ 75
3.5  Comparisons and Conclusions ......................... 80
3.6  Exercises ........................................ 82

4  Automated Smoothing Techniques  85
4.1  Span by Cross-Validation ............................. 86
4.2  Splines and Automated Smoothing ..................... 89
4.2.1  Estimating Smoothing Through the Likelihood 89
4.2.2  Smoothing Splines and Cross-Validation .......... 90
4.3  Automated Smoothing in Practice .................... 93
4.4  Automated Smoothing Caveats ........................ 105
4.5  Exercises ........................................ 107

5  Additive and Semiparametric Regression Models 109
5.1  Additive Models ................................... 110
5.2  Semiparametric Regression Models .................... 112
5.3  Estimation ....................................... 113
5.3.1  Backfitting .................................. 113
5.4  Inference ........................................ 117
5.5  Examples ........................................ 120
5.5.1  Congressional Elections ........................ 120
5.5.2  Feminist Attitudes ................................ 129
5.6  Discussion ....................................... 133
5.7  Exercises ........................................ 134

6  Generalized Additive Models 137
6.1  Generalized Linear Models ......................... 137
6.2  Estimation of GAMS ............................... 140
6.3  Statistical Inference ................................ 141
6.4  Examples ........................................ 143
6.4.1  Logistic Regression: The Liberal Peace .......... 143
6.4.2  Ordered Logit: Domestic Violence ............... 146
6.4.3  Count Models: Supreme Court Overrides ......... 150
6.4.4  Survival Models: Race Riots .................... 154
6.5  Discussion ....................................... 157
6.6  Exercises ........................................ 157
7 Extensions of the Semiparametric Regression Model 161
   7.1 Mixed Models ........................................... 162
   7.2 Bayesian Smoothing .................................... 166
   7.3 Propensity Score Matching ............................ 170
   7.4 Conclusion ............................................. 175

8 Bootstrapping 177
   8.1 Classical Inference .................................... 178
   8.2 Bootstrapping – An Overview ........................... 178
      8.2.1 Bootstrapping ..................................... 179
      8.2.2 An Example: Bootstrapping the Mean .......... 183
      8.2.3 Bootstrapping Regression Models ............... 185
      8.2.4 An Example: Presidential Elections .......... 187
   8.3 Bootstrapping Nonparametric and Semiparametric
      Regression Models ....................................... 189
      8.3.1 Bootstrapping Nonparametric Fits .......... 189
      8.3.2 Bootstrapping Nonlinearity Tests ............. 190
   8.4 Conclusion ............................................. 192
   8.5 Exercises ............................................... 193

9 Epilogue 195

Appendix: Software 197

Bibliography 203

Author Index 209

Subject Index 211
List of Tables

3.1 AIC values for differing numbers of knots ........................................ 62
4.1 Mean squared error for smoother fits ........................................... 96
4.2 Likelihood ratio test between cubic and smoothing splines .............. 103
5.1 Testing spline fits against linear fits and transformations: Challenger’s vote share ................................................................................... 123
5.2 A comparison of parametric and semiparametric models of the 1992 House elections ............................................................. 125
6.1 A comparison of parametric and semiparametric models of Supreme Court overrides .............................................................. 151
6.2 Grambsch and Therneau nonproportionality tests for time to onset of racial riots, 410 US Cities, 1961–1968 ........................................ 155
7.1 GLM and GAM balance test comparison wage data .......................... 173
7.2 GLM and GAM balance test comparison presidential visit data ........ 175
8.1 Hypothetical data on party seat shares ............................................ 183
8.2 Incumbent vote share: Standard and bootstrapped models .............. 188
List of Figures

1.1 A linear fit and the true quadratic relationship with simulated data. . . 5
1.2 Residuals plotted against fitted values with a nonparametric regression estimate. ........................................... 10
2.1 A scatterplot of challenger’s vote share and Perot vote, 1992. . . . . . 18
2.2 A moving average smoother and an OLS regression line. ............ 19
2.3 A moving average smoother: varying the bin width of the smoother. . 20
2.4 An example bin for the Congressional vote data. ..................... 24
2.5 The tricube weighted average. ........................................ 24
2.6 Kernel smoother with three different bandwidths..................... 25
2.7 Local linear regression: fitted $Y$ at the focal point. ................. 27
2.8 Loess and lowess fit to challenger’s vote share and support for Perot, 1992................................................................. 30
2.9 The bias variance tradeoff in smoothing................................ 33
2.10 Selecting the span for a lowess fit to the data....................... 35
2.11 Varying the polynomial degree for local polynomial regression.... 37
2.12 Overfitting a nonparametric regression estimate.................... 38
2.13 95% confidence bands for the loess fit to the challenger’s vote share data................................................................. 43
2.14 The joint effect of overdrafts and challenger spending on the challenger’s vote share.................................................. 46
3.1 A simulated nonlinear functional form with OLS fit.................... 51
3.2 A piecewise linear regression spline fit................................ 54
3.3 Support for the challenger as a function of support for Perot in the 1992 House elections. ........................................ 58
3.4 Cubic B-Spline and natural spline fit to challenger’s vote share. ...... 61
3.5 Differing numbers of knots with natural splines. ..................... 63
3.6 Smoothing spline fit to challenger vote share data with four different degrees of freedom values. ................................. 68
3.7 Smoothing spline fit to challenger vote share data with constant degrees of freedom and different knot numbers. ....................... 69
3.8 Confidence bands for natural cubic spline fit to challenger’s vote share. 77
3.9 A comparison of pointwise variability bands and bias adjusted confidence intervals. .................................................. 78
3.10 Plot of first derivative for the effect of support for Perot on the challenger’s vote share. ................................................. 80
3.11 Comparison of smoother fits. Dotted line is true functional form, solid line is nonparametric estimate. ............................... 81
4.1 GCV plot for local polynomial regression span selection............. 88
4.2 Lowess fit with span selection via GCV. ................................. 89
4.3 Smoothing spline fits with automatic smoothing selection .......... 92
4.4 A comparison of estimated smoother fits. ................................ 95
4.5 A simulated nonlinear smooth functional form. .......................... 97
4.6 Comparing fits between smoothing splines and lowess with various spans. ................................................................. 98
4.7 Lowess span chosen by MSE and smoothing spline fit. ............ 99
4.8 Manual selection of degrees of freedom compared to automatic selection for overdrafts and challenger’s vote share. ............... 100
4.9 Scatterplot of county level votes for Al Gore and overvotes with smoothing spline fits, Florida 2000. ............................... 102
4.10 Standardized residual plots with smoothing spline fits. .......... 106
5.1 Additive model for challenger’s vote share as a function of support for Perot and the number of overdrafts. .......................... 121
5.2 Nonparametric estimates for semiparametric regression model of Congressional challenger’s vote share, 1992. ....................... 126
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3</td>
<td>Interaction between support for Perot and competitive districts.</td>
<td>128</td>
</tr>
<tr>
<td>5.4</td>
<td>Interaction between challenger spending and the number of overdrafts.</td>
<td>129</td>
</tr>
<tr>
<td>5.5</td>
<td>The effect of income proportion, hours worked, and age on feminist attitudes, women (1987–2000).</td>
<td>131</td>
</tr>
<tr>
<td>5.6</td>
<td>The effect of income proportion on feminist attitudes with manual smoothing selection.</td>
<td>132</td>
</tr>
<tr>
<td>6.1</td>
<td>The nonparametric estimates for risk of a militarized dispute.</td>
<td>144</td>
</tr>
<tr>
<td>6.2</td>
<td>Comparison of the effect of democracy on militarized conflict: probability scale.</td>
<td>145</td>
</tr>
<tr>
<td>6.3</td>
<td>The effect of relationship duration on couple violence: linear predictor scale.</td>
<td>147</td>
</tr>
<tr>
<td>6.4</td>
<td>The effect of relationship duration on couple violence: probability scale.</td>
<td>148</td>
</tr>
<tr>
<td>6.5</td>
<td>Density plot for relationship duration.</td>
<td>149</td>
</tr>
<tr>
<td>6.6</td>
<td>The difference in the effect of tenure across parametric and semiparametric models.</td>
<td>152</td>
</tr>
<tr>
<td>6.7</td>
<td>Nonparametric estimate for Congressional counter variable.</td>
<td>153</td>
</tr>
<tr>
<td>6.8</td>
<td>The effect of manufacturing wages and non-white unemployment on the hazard of racial riots, 1961–1968.</td>
<td>156</td>
</tr>
<tr>
<td>7.1</td>
<td>The effect of family SES on math achievement for parochial and public schools.</td>
<td>165</td>
</tr>
<tr>
<td>7.2</td>
<td>Comparison of Bayesian and frequentist estimates of challenger spending on challenger’s vote share.</td>
<td>169</td>
</tr>
<tr>
<td>8.1</td>
<td>Q-Q plot of the bootstrap sampling distribution.</td>
<td>184</td>
</tr>
<tr>
<td>8.2</td>
<td>Histogram and Q-Q plot of bootstrap sampling distribution for policy moderation.</td>
<td>187</td>
</tr>
<tr>
<td>8.3</td>
<td>Plug-in and bootstrap confidence intervals for effect of Perot vote on challenger vote share.</td>
<td>190</td>
</tr>
</tbody>
</table>
Preface

This is a book for analysts from the social sciences who would like to extend their statistical toolkit beyond standard parametric functional forms. In general, the methods presented in this book alter the use of standard models. That is, the techniques here are not meant to be a replacement for models such as linear or logistic regression, but instead are meant to enhance how these models are used. As a result, the techniques can enhance data analysis but they are not a panacea. While nonparametric regression can flexibly estimate nonlinear functional forms, it cannot make a correlation into a causal effect, it cannot make up for an omitted variable, and it cannot prevent data mining. The visual aspect of these models, however, can make analysts more sensitive to patterns in the data that are often obscured by simply reading a parameter estimate off the computer screen.

The goal of this book is to make applied analysts more sensitive to which functional form they adopt for a statistical model. The linear functional form is ubiquitous in social science research, and it often provides a very reasonable approximation to the process we are trying to understand. But it can also be wrong, leading to inferential errors, especially when applied without much thought as to whether it is justified. We do not presume that all relationships are nonlinear, but that the assumption of linearity should be tested, since nonparametric regression provides a simple and powerful tool for the diagnosis and modeling of nonlinearity.

This book is also designed to be attuned to the concerns of researchers in the social sciences. There are many books on this topic, but they tend to focus on the concerns of researchers in the sciences and are full of examples based on the patterns of sole eggs and the reflection of light off gas particles. To that end, we provide a hands-on introduction to nonparametric and semiparametric regression techniques. The book uses a number of concrete examples, many from published research in both political science and sociology, and each example is meant to demonstrate how using nonparametric regression models can alter
conclusions based on standard parametric models. The datasets and R code for every example can be found on the book webpage at: http://www.wiley.com/go/keele_semiparametric. Moreover, the appendix provides a basic overview of how to produce basic results.

Since the techniques presented here alter the estimation of standard models, we assume that the reader is familiar with more standard models. For example, while we provide an example of estimating a semiparametric count model, we do not outline aspects of count models. Instead, we refer readers to more authoritative sources. In general, various chapters have differing levels of presumed knowledge. Chapters 1–3 presume a basic familiarity with linear regression models at a level that would be similar to a course in regression in a graduate first-year methods sequence. That said, various topics within those chapters may be more advanced. If this is the case, we alert readers. Chapter 6 presumes that readers are familiar with more advanced models including discrete choice models such as logit and probit, count models, and survival models. The presentation of these models, however, is self-contained such that readers familiar with logit models can read the relevant sections without needing to know anything about survival models. The more advanced topics in Chapter 7 again each presume specialized background information but are again self-contained.

I would like to thank the many people who provided feedback and helpful comments including Chris Zorn, Alan Wiseman, Irfan Nooruddin, Marcus Kurtz, Charles Franklin, Karen Long Jusko, Michael Colaresi, Mike Neblo, Craig Volden, Jan Box-Steffensmeier, Neal Beck, Jas Sekhon, Anand Sokhey, Ryan Kennedy, Corrine McConnaughy, Mark Kayser, and the many anonymous reviewers. I am also indebted to all the people who provided me with data for use in the illustrations, and to my students for allowing me to try the material out on them. I also want to acknowledge Jim Stimson and Walter Mebane for general guidance and advice on what it means to do careful research using data. Finally, I would like to thank my wife Tracey for not begrudging all the nights and weekends spent with data instead of her. The author would appreciate being informed of any errors and may be contacted at: keele.4@osu.edu.

Luke Keele
1

Introduction: Global versus Local Statistics

Statistical models are always simplifications, and even the most complicated model will be a pale imitation of reality. Given this fact, it might seem a futile effort to estimate statistical models, but George Box succinctly described the nature of statistical research: ‘All models are wrong, some are useful.’ Despite the fact that our models are always wrong, statistics provides us with considerable insight into the political, economic and sociological world that we inhabit. Statistical models become simplifications of reality because we must make assumptions about various aspects of reality. The practice of statistics is not shy about the need to make assumptions. A large part of statistical modeling is performing diagnostic tests to ensure that the assumptions of the model are satisfied. And in the social sciences, a great deal of time is devoted to checking assumptions about the nature of the error term: are the errors heteroskedastic or serially correlated? Social scientists are far more lax, however, when it comes to testing assumptions about the functional form of the model. In the social sciences, the linear functional (and usually additive) form reigns supreme, and researchers often do little to verify the linearity assumption. Much effort is devoted to specification and avoiding misspecification, but little is done to explore other functional forms when the incorrect functional form is in essence a specification error.
The reliance on linear functional forms is more widespread than many probably realize. For many analysts, the question of linearity is focused solely on the nature of the outcome variable in a statistical model. If the outcome variable is continuous, we can usually estimate a linear regression model using least squares. For discrete outcomes, analysts typically estimate generalized regression models such as logistic or Poisson regression. Researchers often believe that because they are estimating a logistic or Poisson regression model, they have abandoned the linear functional form. Often this is not the case, as the functional form for these models remains linear in an important way. The generalized linear model (GLM) notation developed by McCullagh and Nelder (1989) helps clarify the linearity assumption in models that many researchers think of as nonlinear.

In the GLM framework, the analyst makes three choices to specify the statistical model. First, the analyst chooses the stochastic component of the model by selecting a sampling distribution for the dependent variable. For example, we might choose the following sampling distribution for a continuous outcome:

$$Y_i \sim N(\mu_i, \sigma^2).$$  \hspace{1cm} (1.1)

Here, the outcome $Y_i$ follows a Normal sampling distribution with expected value $\mu_i$ and a constant variance of $\sigma^2$. This forms the stochastic component of the statistical model. Next, the analyst must define the systematic part of the model by choosing a set of predictor variables and a functional form. If we have data on $k$ predictors, $\mathbf{X}$ is an $n \times k$ matrix containing $k$ predictors for $n$ observations, and $\eta$ is an $n \times 1$ vector of linear predictions:

$$\eta = \mathbf{X}' \mathbf{\beta}.$$  \hspace{1cm} (1.2)

where $\mathbf{\beta}$ is a vector of parameters whose values are unknown and must be estimated. Both $\mathbf{X}' \mathbf{\beta}$ and $\eta$ are interchangeably referred to as the linear predictor. The linear predictor forms the systematic component of the model. The systematic component need not have a linear functional form, but linearity is typically assumed. Finally, the analyst must choose a link function. The link can be as simple as the identity function:

$$\mu_i = \eta.$$  \hspace{1cm} (1.3)

The link function defines the connection between the stochastic and systematic parts of the model. To make the notation more general, we can write the link function in the following form:

$$\mu_i = g(\eta_i).$$  \hspace{1cm} (1.4)

where $g(\cdot)$ is a link function that must be monotonic and differentiable. When the stochastic component follows a Normal distribution and the identity function links the stochastic and systematic components, this notation describes a
linear regression model. The GLM framework, however, generalizes beyond linear regression models. The stochastic component may come from any of the exponential family of distributions, and any link function that is monotonic and differentiable is acceptable.

As an example of this generality, consider the GLM notation for a model with a binary dependent variable. Let $Y_i$ be a binary variable without outcomes 0/1, where 1 represents a success for each $m_i$ trials and $\psi_i$ is the probability of a success for each trial. If this is true, we might assume that $Y_i$ follows a Bernoulli distribution, which would imply the following stochastic component for the model:

$$Y_i \mid \psi_i \sim B(m_i, \psi_i).$$

(1.5)

The systematic component remains $X'\beta$, the linear predictor. We must now select a link function that will ensure that the predictions from the systematic component lie between 0 and 1. The logistic link function is a common choice

$$\psi_i = \frac{1}{1 + e^{-\eta_i}}.$$  

(1.6)

With the logit link function, no matter what value $\eta_i$ takes the predicted value for $Y_i, \psi_i$, will always be between zero and one.

What has been the purpose of recounting the GLM notation? The point of this exercise is to demonstrate that while the link function is a nonlinear transformation of the linear predictor, the systematic component of the model remains linear. Across the two examples, both the link function and stochastic component of the model differed, but in both cases we used the linear predictor $X'\beta$. The second model, a logistic regression, is thought of as a nonlinear model, but it has a linear functional form. Thus many of the models that analysts assume are nonlinear models retain a linearity assumption. There is nothing about the logistic regression model (or any other GLM) that precludes the possibility that the model is nonlinear in the variables. That is, instead of the effect of $X$ on $Y$, which is summarized by the estimated $\beta$ coefficient, being constant, that effect varies across the values of $X$. For example, one is as likely to use a quadratic term in a Poisson regression model as a linear regression model.

Why, then, is the assumption of linearity so widespread? And why is model checking for deviations from nonlinearity so rare? This question cannot be answered definitively. Quite possibly, the answer lies in the nature of social science theory. As Beck and Jackman (1998) note: ‘few social scientific theories offer any guidance as to functional form whatsoever.’ When researchers stipulate the predicted relationship between $X$ and $Y$, they do not go beyond ‘the relationship between $X$ and $Y$ is expected to be positive (or negative).’ Given that most
theory is silent as to the exact functional form, linearity has become a default, while other options are rarely explored. We might ask what are the consequences of ignoring nonlinearity?

1.1 The Consequences of Ignoring Nonlinearity

What are the consequences for using the wrong functional form? Simulation provides some useful insights into how misspecification of functional form can affect our inferences. First, we need some basic notation. We start with the linear and additive functional form for \( Y \)

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \tag{1.7}
\]

The above equation is linear in the parameters in that we have specified that the mean of \( Y \) is a linear function of the variables \( X_1 \) and \( X_2 \). It is possible that the effect of \( X_2 \) on \( Y \) is nonlinear. If so, the effect of \( X_2 \) on \( Y \) will vary across the values of \( X_2 \). Such a model is said to be nonlinear in the variables but linear in the parameters. Estimation of models that are nonlinear in the variables presents few problems, which is less true of models that are nonlinear in the parameters. Consider a model of this type

\[
Y = \beta_0 + \beta_1 X_1^{\beta_1}. \tag{1.8}
\]

Here, the model is nonlinear in the parameters and estimation is less easy as nonlinear least squares is required. For a model that is nonlinear in the variables, we can still use least squares to estimate the model. What are the consequences of ignoring nonlinearity in the variables if it is present? For example, assume that the true data generating process is

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2. \tag{1.8}
\]

Suppose an analyst omits the quadratic term when estimating the model. The effect of omitting the nonlinear term from the right hand side of a regression model is easily captured with simulated data. To capture the above data generating process, we simulate \( X_1 \) as 500 draws from a uniform distribution on the interval 1 to 50. \( Y \) is a function of this \( X \) variable with an error term drawn from a Normal distribution with zero mean and constant variance and the \( \beta \) parameters are set to ones. What are the consequences of fitting a linear model without the quadratic term? Figure 1.1 contains a plot of the estimated regression line when only a linear term has been included on the right hand side of the estimated model and the true quadratic functional form.
Here, the model with the incorrect functional form would lead one to conclude that $X$ is unrelated to $Y$ as the regression line is virtually flat, when in fact the curved line reflects the strong but nonlinear relationship between the two variables. This simulated example illustrates the misspecification that results from assuming linearity when the relationship between $X$ and $Y$ is actually nonlinear. In this particular example, the consequences are particularly severe, given that an analyst might conclude that there is no relationship between $X$ and $Y$, when the two are strongly related. Moreover, as the GLM framework implies, models such as logistic or Poisson regression are equally prone to this misspecification from an incorrect functional form. If $X^2$ is in the true data generating process, no link function will correct the specification error. In these models, the systematic component of the model often remains linear and the failure to include a nonlinear term in the model will have equally deleterious effects on the model estimates. In fact, given the serious distortions such a misspecification can cause, it would be prudent to test that the effect of any continuous covariate on the right hand side of a model does not have a nonlinear effect. If we admit that defaulting to linearity might be a problematic method of data analysis, what alternatives are available? While there are a number of possibilities, analysts in the social sciences usually rely on power transformations to address nonlinearity.

1.2 Power Transformations

Power transformations are a simple and flexible means of estimating a nonlinear functional form. While a variety of power transformations are possible, most
researchers restrict themselves to only one or two transformations. A common notation exists for all power transformations that is useful to outline since we will often use power transformations in the data analytic examples that follow. For a strictly positive variable, $X$, we can define the following set of power transformations:

$$X^\lambda.$$  \hfill (1.9)

Using different values for $\lambda$ produces a wide variety of transformations. If $\lambda$ is 2 or 3, the transformation is quadratic or cubic respectively, and if $\lambda$ is $1/2$ or $1/3$, the transformation is either the squared or cubic root. By convention, a value of 0 for $\lambda$ denotes a log transformation, and a value of 1 for $\lambda$ corresponds to no transformation (Weisberg 2005).

Power transformations are often a reasonable method for modeling nonlinear functional forms. For example, it is well understood that older people vote more often than younger people (Nagler 1991). So as age increases, we expect that the probability of a person voting increases. But we do not expect the effect of age on voter turnout to be linear since once people reach a more advanced age, it often prevents them from voting. To capture such nonlinearity, most analysts rely on a quadratic power transformation. In this model, the analyst would square the age variable and then include both the untransformed and squared variable on the right hand side of the model. If the quadratic term is statistically significant at conventional levels ($p < 0.05$), the analyst concludes that the relationship between the two variables is nonlinear.

The process described above is often a reasonable way to proceed. Some care must be taken with the interpretation of the quadratic term as the standard error for the marginal effect must be calculated by the analyst, but the transformation method is easily to use. Moreover, the model is now a better representation of the theory, and power transformations avoid any misspecification due to an incorrect functional form. If the effect of age on voter turnout is truly quadratic, and we only include a linear term for age; we have misspecified the model and biased not only the estimate of age on voter turnout but all the other estimates in the model as well.\footnote{This is true since I assume the analyst has estimated a model such as a logit. In linear regression, misspecification only biases the coefficient estimates if the omitted variable is correlated with other variables, but for most models with nonlinear link functions, all the parameters will be biased if the model is misspecified.}

Power transformations, however, have several serious limitations. First and foremost, power transformations are global and not local fits. With a global fit, one assumes that the statistical relationship between $X$ and $Y$ does not vary over the range of $X$. When an analyst estimates a linear relationship or uses a